When Machines Evaluate Machines.

An empirical study on the impact of chunk size and top_k, and the reliability of LLM-based evaluation.

Darpan Beri

GROUP
A-20

AI’s Evaluation of other AI’s is Only Moderately Reliable!

Less is More when providing Information to AI!

Introduction

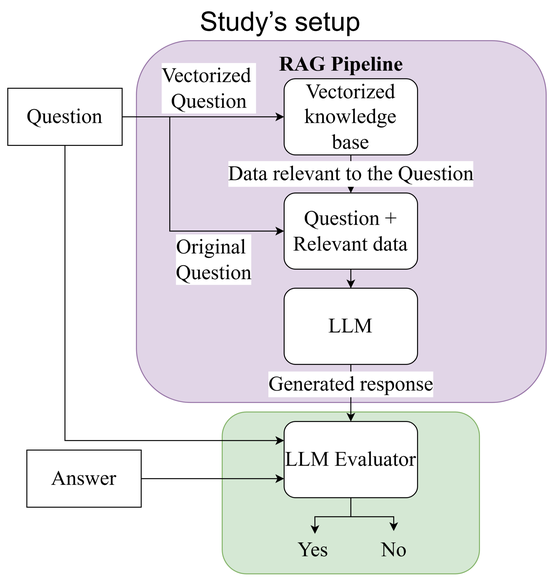
LLMs are revolutionary yet struggle with knowledge cutoffs and hallucinations. This study answers two questions at:

- How do the settings of the LLM pipeline impact the response?
- How reliable is an LLM-based evaluator as compared to human judgement?

Chunk_size (number of words in a paragraph)	Top_k (number of paragraphs)	Example Chunks (Visible to LLM within the RAG model)
50	2	Chunk 1 (50 words): “The mitochondrion is the powerhouse...” Chunk 2 (50 words): “ATP is the primary energy currency of the cell...”

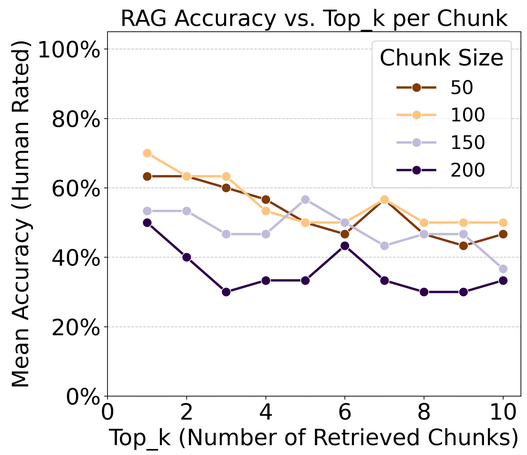
Method

We asked Retrieval-Augmented Generation (RAG) to programmatically evaluate LLMs as follows:



Results

Best accuracy (~70%) for the RAG model achieved with chunksize = 100 and top_k = 1.



Best accuracy (~70%) for the RAG model achieved with chunksize = 100 and top_k = 1. LLM evaluator was moderately accurate (73.4%, Cohen’s Kappa = 0.47). Providing excessive information to the RAG reduced accuracy.